# Fahad – Ideas so far (Thursday, June 26, 2014)

## First write up

*"It is not worth asking how to define consciousness, how to explain it, how it evolved, what its function is, etc., because there's not one thing for which all the answers would be the same. Instead, we have many sub-capabilities, for which the answers are different: e.g., different kinds of perception, learning, knowledge, attention, control, self-monitoring, self-control, etc."* --Aaron Sloman (1994)[1]

I'd like to look at this project through the lens of Computer Science and Artificial Intelligence, i.e., how would one imbue an artificial agent (software agent, robot, etc.) with consciousness? My primary source of information is Marvin Minsky's book *"The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind."* [2]. In reference to consciousness, the most important message in this book is that we should realize that consciousness is not just one thing (as is the message in Sloman's quote above). It is a "suitcase word" (as Minsky calls it) that is used to describe several different things. The breakup of consciousness into many parts is available here [3].

The most important aspects of conscious experience to me are Emotions (especially Compassion, Empathy, etc), Self-Awareness, Reflection and Self-Reflection (possibly a few more). I find these phenomena particularly interesting because unlike learning and planning (which we know how to program), how to design the previously mentioned phenomena for artificial agents is not commonly known outside the field of AI. The idea is to think of each of these phenomena as emerging from interactions between micro-agents (e.g., Neurons, Glia, etc.) through various layers of hierarchy. The hierarchical system structure is a fairly common design for intelligent software agents (e.g., simulated ice hockey players) and robots. Such a hierarchical framework, I believe, simplifies design and simplifies our perspective on these complex phenomena.

Think about how a robot would know where it is, i.e., imagine a robot asking itself, "where am I?". This is the problem of "Mobile Robot Localization". We don't even think about having that problem. But for a robot, it is indeed quite a challenging problem. Nevertheless, we now have algorithms that make it possible for mobile robots to navigate rough terrains. But now imagine the robot asking, "who am I"? "why do I exist?" ... those are the questions, in my opinion, the answers to which require the phenomena we aggregate under the wide umbrella of consciousness. Oh, by the way, don't forget how important a role language and expression plays in our conscious experiences as humans.

To summarize, I'd like to look into the following question: "How does one program a software agent that wonders about its identity, and why it exists? ...". It would be interesting to see what small modules/tasks (analogous to Neurons/Glia) need to be integrated together to make that happen. Perhaps that will provide us with some insight into the emergence of consciousness?

## Second Write up

Luckily, I found a review paper that describes what the most prominent AI researchers think about consciousness. Here's the paper[7]. Summary: "The idea in a nutshell is that *phenomenal consciousness* is the property a computational system X has if X models itself as experiencing things."

## The "Mind -- Software" analogy

**Hypothesis:** Mind emerges from the Brain, i.e., the interactions between neurons and glia. We know that a neural network is a universal computer, and the assumption here is that the computation is expressed as a neural network. So we look only at this level, and assume that the underlying molecular mechanisms provide the medium for computation. *Explanation:* Think of the brain as hardware made out of Neurons, Glia, and other stuff. Essentially, the brain is a *network*. The network *structure* is how the components are interconnected. *Dynamics* on this network are defined as the interactions between the massive number of components. Mind (and as a result consciousness) is a *function* of this network. So the brain is a network that does computation.

In a computer, the hardware is the microprocessor, memory, etc. All the high-level software output that we see as, e.g., animated Minions, is an emergent phenomenon of the programmed interaction between hardware components.

Brain ←→ Hardware :: Mind ←→ Software

**Simulation of Consciousness**

As mentioned in the earlier Wiki post, simulation of consciousness can be achieved by using Minsky's breakdown. Tentative plan:

- Implement the various properties of consciousness in agents
- Use a simplified model of language, which can be used by the agents to express their thoughts, ideas, etc.
- It seems like there are lots of things about consciousness which we only perceive because we can either observe them expressed in language, or facial expression, etc.
- If we really want to see whether the agents are thinking or not, and what emotions they have, we have to have multiple agents that learn different things and should therefore have different sorts of emotions; even for similar situations perhaps.
- We can then compare the different situations and emotions to see if there is enough evidence for consciousness.
- Try to come up with a model that is as simple as possible

**Potential Issues**

- *Natural language processing is a hard computational problem.* It may not be possible to get as much out of a simulation as we'd like, simply because we run into issues of compute power.
- *Gödel's incompleteness Theorems*. There are truths that we cannot prove mathematically. Perhaps there is something about consciousness that is inherently quantum mechanical, and cannot be expressed using classical computation.

**My Perspective on John's Consciousness Table**

| | |
|---|---|
| Requires matter | Yes |
| Exclusive property of biological systems | Yes |
| Has parts | Yes |
| Admits of gradations or degree (continuous or discrete) | Yes |
| Likely to be artificial created in the AI lab | Yes |
| Intrinsically paradoxical research area. (involves strange loopy-ness Gödel, Turing) | Yes |
| Requires interaction with external environment (perception?) | Yes |
| Requires interaction with internal environment (reflection?) | Yes |
| Necessarily Embodied (result of sensor interactions, hormonal environment etc.) | Yes |
| Necessarily Embedded (result of large scale system interactions) | Yes |
| Deterministic system with stochastic inputs | ? |
| Relies on quantum things that I do/do not understand | ? |
| Defined by its ineffability, will always recede from scientific apprehension | No |
| Encodes history (requires memory, genes) | Yes |
| Is multifunctional? | ? |
| Ever changing? | Yes |
| Life-contingent? | ? |